



# Audio-Based Classification and Geographic Regression of Austrian Dialects

Lorenz Gutscher<sup>1,2</sup>, Michael Pucher<sup>1,2</sup>

<sup>1</sup>Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

<sup>2</sup>Austrian Research Institute for Artificial Intelligence, Vienna, Austria

lorenz.gutscher@ofai.at, michael.pucher@ofai.at

## Abstract

Dialect classification remains challenging due to regional variability and limited dialect-specific datasets. This study addresses these challenges by leveraging a novel dataset of 304 speakers from 108 locations across Austria for automatic classification of Austrian dialects. To minimize speaker-specific biases and enhance dialectal features, speaker augmentation techniques are applied. Classification is conducted at three levels: location, dialect group, and federal state. Additionally, a regression task predicts the speakers' geographic coordinates, with the wav2vec 2.0 model architecture achieving an average test-set distance error of 66.7 kilometers. This work represents a unique approach to fine-grained dialect classification and geographic location prediction for Austria. Finally, model explainability is explored using Integrated Gradients (IG), identifying the most relevant speech segments for classification within each dialect group.

**Index Terms:** Dialect classification, Austrian dialects, explainable AI, data augmentation

## 1. Introduction

The two primary dialect families spoken in Austria are Alemannic (Alemannisch) and Bavarian (Bairisch). A more fine-grained classification further divides these into six subgroups: Alemannic (Alem.), Bavarian-Alemannic (Bav.-Alem.), South Bavarian (S. Bav.), South-/Central Bavarian (S./C. Bav.), West-Central Bavarian (W.-C. Bav.), and Central Bavarian (C. Bav.), as illustrated in Figure 1. These dialects differ in phonetics, vocabulary, and grammar, reflecting the linguistic diversity shaped by historical and geographical influences [1, 2]. In addition to dialects, Standard Austrian German is spoken all over Austria and can be seen as a variety situated between the C. Bav. and Standard German spoken in Northern Germany [3].

Dialect classification enhances speech technology systems, such as Automatic Speech Recognition (ASR), by incorporating dialect embeddings – high-dimensional feature representations extracted from audio – shown to improve word error rate [4]. This is especially relevant for Austrian dialects, where dialectal variation in conversational spoken input is a challenge for state-of-the-art speech recognition architectures [5]. Key challenges include dialectal variation within dialect groups and transitional regions, where dialect boundaries are fluid and overlapping. While dialect classification can also be performed at the lexical level [6], this work focuses on phonetic characteristics.

Dialect groups and federal states can be classified either directly or by training on speakers' places of origin, enabling evaluation at the location level and comparison with direct classification models. If locations are used as geographic coordinates, the classification task can be reformulated as a regression task

predicting latitude and longitude.

To gain insight into model decisions, we apply Integrated Gradients (IG) to identify speech segments that contribute most to the classification. This post hoc explainability offers valuable information for Austrian dialect speakers and linguists, helping to uncover the linguistic cues the model relies on. By making Artificial Intelligence (AI) decisions more transparent, explainability can increase trust and acceptance of such systems [7].

This paper makes the following contributions:

- Classification of location, dialect group, and federal state
- Predicting geographic coordinates in a regression task
- Evaluation of speaker augmentation
- Explainability using attribution analysis

## 2. Related Work

Most prior work on dialect classification from audio has focused on accent classification, predominantly for English, as demonstrated in [8] and [9]. In contrast, fine-grained dialect classification – particularly for Austrian dialects – has received comparatively less attention. While classification of German dialect families has been addressed in [4] and [10], Austrian data in these works is grouped coarsely into Alemannic and Austrian-Bavarian, limiting specificity. Another study focuses exclusively on Austrian dialects but limits its scope to the two most linguistically distinct federal states, Vienna (Vie) and Vorarlberg (Vbg), achieving an accuracy of 84.2% [11]. Beyond classification, statistical analyses of Austrian dialects have been conducted, examining phonological variable patterns derived from individual tokens and isolated linguistic features [12]. Geographic regression approaches, such as [13], have also been applied to dialectal variation, utilizing random forests to model German phone classes.

In the field of explainability, Layer-wise Relevance Propagation has been used to analyze feature importance on raw waveforms in gender and digit classification [14]. In [15], explainable AI techniques are applied to speech-based Alzheimer's screening. A Non-negative Matrix Factorization method is proposed in [16] to enhance interpretability in scene classification, which, though originally applied to audio events, also holds potential for dialect classification. Source separation methods such as the one proposed in [17] require prior source knowledge, limiting their use in dialect classification.

## 3. Methodology

This section presents the dataset – comprising dialect speech recordings from selected Austrian locations – along with the data augmentation techniques, modeling approaches, and explainability methods applied in this study. Classification and

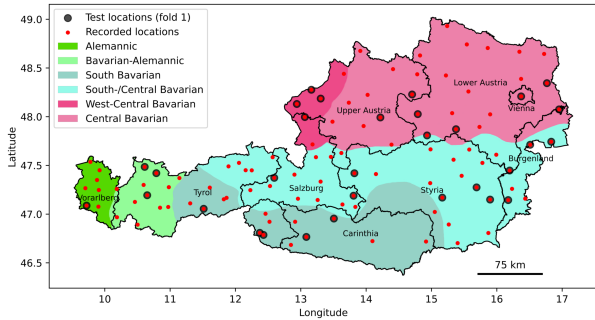


Figure 1: Map of Austria showing recording locations and their associated dialect groups. Adapted from [12] and [22].

regression models use ECAPA-TDNN [18] and wav2vec 2.0-based architectures [19, 20] to predict dialect group, federal state, and location classes, as well as geographic coordinates via regression. Additionally, IGs [21] are employed to provide explainability of the model and its decision-making process, proposing two methods: (1) Identifying phones at attribution peaks; (2) Measuring kurtosis to identify utterances with distinctive phones.

### 3.1. Dataset

In the case of Austria, corpora like Common Voice [23] lack precise dialect-specific metadata, which is essential for fine-grained dialect classification. Therefore, the corpus used in this study is derived from the following sources: (1) A corpus collected as part of the research program *German in Austria: Variation and Change of Dialect Varieties in Austria (in Real and Apparent Time)*<sup>1</sup> which consists of recordings from 106 locations. (2) Two new locations, Vie [24] and Innervillgraten, are added, and four speakers from Bad Goisern (which is also listed in (1)) are included [25]. All data are downsampled to 16 kHz. The combined dataset comprises 304 speakers from 108 regions in Austria, selected based on criteria including low formal educational attainment, engagement in manual occupations (e.g., agriculture), and long-term residence in the same area. It includes 153 elderly speakers (65+ years: 69 female, 84 male) and 151 young speakers (18–35 years: 66 female, 85 male). For each region, there are at least two and at most four individuals. The recordings consist of dialogues between the investigator and the speaker, with the investigator asking a fixed set of questions. These include open-ended questions, translation tasks, and structured prompts aimed at eliciting target phones for later analysis. Using automatic speaker diarization [26], the dataset is processed to remove the investigator, ensuring that only the dialect speakers are present in the data. Long utterances are segmented into 5-second chunks, and the minimum duration of an utterance is set to one second. This results in a dataset with an average of 612 utterances per speaker and a mean duration of 2.02 seconds per utterance. The dataset is imbalanced across Austria’s nine federal states, with the number of speakers ranging from three speakers in Vie to 64 in Tyrol (Tyr). The distribution across the other states is as follows: Burgenland (Bgl, 14), Carinthia (Car, 22), Lower Austria (L. Aut, 48), Upper Austria (U. Aut, 49), Salzburg (Sbg, 38), Styria (Sty, 42), and Vbg (24). This imbalance is also reflected in the num-

<sup>1</sup><https://www.dioe.at/projekte/task-cluster-b-variation/pp02>

ber of utterances, with Vie having the fewest (1,506) and Tyr the most (40,098). Similarly, dialect groups also exhibit an imbalance in speaker and utterance counts, as shown in Table 1.<sup>2</sup>

Table 1: Number of speakers and utterances per dialect group

Dialect group	Speakers	Utterances
Alemannic	28	15,884
Bavarian-Alemannic	24	16,141
South Bavarian	48	28,967
South-/Central Bavarian	102	60,468
West-Central Bavarian	28	17,847
Central Bavarian	74	46,804

### 3.2. Data augmentation

Data augmentation is widely used in speech tasks to improve robustness, especially when training data lacks sufficient variability [27, 28]. Large foundation models no longer require data augmentation [29]; however, to achieve dialect robustness with limited resources, it remains an important technique. Each speaker’s recordings are converted into 50 voice variants using reference speakers, while preserving dialectal features, as proposed in [30], ensuring high-quality recordings without requiring transcriptions. Reference speakers are manually selected from the Common Voice dataset, specifically Standard Austrian German speakers with perceptually clean recordings (25 female, 25 male). The voice-converted samples aim to match the source dialect while preserving dialect-specific features, though subtle pronunciation changes may still occur, slightly altering the original dialect. By standardizing speaker characteristics (timbre) across dialects through voice conversion, the model is encouraged to focus on dialect- or language-specific attributes rather than speaker-specific attributes. Additionally, added noise and speed perturbations (0.95, 1.05) are tested for their ability to improve accuracy by increasing speaker diversity, though they may affect subtle dialectal cues due to pitch shifts.

### 3.3. Classification and regression task

For model training, ECAPA-TDNN [18], a speaker verification model designed for robust feature extraction from speech, is employed and serves as a baseline. While originally developed for speaker recognition, it can also be fine-tuned for tasks such as dialect classification due to its ability to capture speaker-independent features. Additionally, the self-supervised wav2vec 2.0 architecture, a transformer-based model, is used. Wav2vec 2.0 has demonstrated strong performance in learning speech representations for ASR, leading to the development of XLSR-53 (hereafter referred to as XLSR) [19], which has been fine-tuned on multilingual speech data to enhance its effectiveness for tasks like ASR and language classification. The Massively Multilingual Speech (MMS) model [20] is based on the wav2vec 2.0 architecture and trained on a significantly larger dataset spanning over 4,000 languages. A fine-tuned MMS variant (MMS-LID 256, further referenced as MMS-LID) for lan-

<sup>2</sup>Once the raw dataset is published by the project group, the processed dataset (with separated/augmented speakers) is expected to be released and could serve as a valuable resource for spoken Austrian dialects.

guage identification covering 256 languages – including German – is selected for this study. This choice ensures coverage of languages potentially related to Austrian dialects while maintaining focus on diverse linguistic representations. Our implementation builds upon [9]<sup>3</sup>. To obtain a single vector representation of dialect embeddings per utterance, all models employ a `StatPooling()` layer, as described in [9]. For classification tasks, the negative log-likelihood loss is used, while mean squared error loss is used for regression.

Despite the expected low accuracy due to the limited number of speakers per location, classification at the location level enables evaluation at the dialect group and state levels. However, such classification does not explicitly capture dialect relationships. To address this, locations can be converted to geographic coordinates, reformulating the task as a regression problem that predicts geographic latitude and longitude. It should be noted that strict dialect boundaries may not be accurately captured in a regression approach, as regression inherently relies on interpolation and does not account for discrete linguistic borders. Furthermore, the regression model cannot capture a mixture of dialect features present in the test data, i.e., from a speaker that has acquired different dialect competencies. For such a scenario, regression and classification models need to be combined.

### 3.4. Explainability

To gain a deeper understanding of the model’s classification criteria, a post-hoc analysis is conducted to assess the IG of the audio. This is achieved using the method from the Captum package [31]. The primary objective of this explainability approach is to identify the most relevant time intervals and phones in the audio that contribute to the model’s classification decisions. Attributions are examined at different stages of the model, including the raw waveform, extracted features, and deeper layers. The tests demonstrated that the attributions derived from the raw waveform provided the most promising results for the task at hand. While this method identifies regions with the highest attribution, it may not fully capture intricate temporal patterns. The workflow for the explainability analysis is as follows: IGs are computed using 50 steps, with a baseline set to zeros. The absolute values of the attributions are smoothed using a Savitzky-Golay filter [32] (window size = 511, polynomial order = 3), and the peaks are identified with a minimum time distance of 0.2 seconds. The timestamp of the peak is used to find the corresponding phone in the audio. The phone transcriptions are extracted using a multilingual model<sup>4</sup> [33] with German selected as the target language. Although this model is primarily trained on Standard German, it provides a rough estimate of which phones are most prominent at the peaks. Additionally, kurtosis is used to quantify IG signal spikiness, enabling tests of whether peaks are higher in prompts targeting specific phones.

## 4. Results

This section presents the performance evaluation of all trained models across different prediction tasks, beginning with model naming conventions and training details, followed by results from Table 2, and findings on model explainability. Models are labeled by prediction task: `_loc` (location classification), `_grp` (dialect group classification), `_state` (federal state classification), and `_coord` (coordinate regression). The suffix `_a` indicates mod-

<sup>3</sup><https://github.com/JuanPZuluaga/accent-recog-slt2022>

<sup>4</sup><https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft>

Table 2: Model performance across different models (fold 1); in case of “`_coord`” and “`_loc`” models, dialect group (Group) and federal state (State) are derived from location (Loc.) results

Model	Distance (km)	Accuracy (%)		
		Loc.	Group	State
mms_coord_a	<b>66.7</b>	<b>10.8</b>	<b>66.3</b>	55.1
mms_coord	69.5	10.1	63.7	<b>56.5</b>
mms_loc	111.7	<b>10.8</b>	50.8	44.9
xlsr_coord	77.8	4.4	51.7	51.2
xlsr_loc	107.9	10.7	52.1	46.2
xlsr_loc_nospkaug	139.3	6.5	32.3	23.8
ecapa_coord	119.9	2.4	38.6	30.7
ecapa_loc	115.9	8.3	48.4	41.0
ecapa_loc_nospkaug	153.3	4.4	31.5	24.3
<b>Trained directly on dialect group or federal state</b>				
mms_grp / mms_state	-	-	65.9	60.1
xlsr_grp / xlsr_state	-	-	71.3	<b>64.3</b>
xlsr_grp_a / xlsr_state_a	-	-	<b>72.7</b>	64.0

els trained with added noise and spectral augmentation. Models based on ECAPA-TDNN, MMS-LID, and XLSR are denoted by the prefixes `ecapa`, `mms`, and `xlsr`, respectively. Models `xlsr_loc_nospkaug` and `ecapa_loc_nospkaug` exclude speaker augmentation, which is applied in all others. Training stops after 25 epochs as the validation error plateaus, indicating diminishing returns. Training time per epoch ranges from five minutes for ECAPA-TDNN ( $\approx 21$  million parameters) to 45 minutes for XLSR ( $\approx 315$  million parameters) and two hours for MMS-LID ( $\approx 968$  million parameters) on an RTX 4090 GPU. The dataset employs 10-fold cross-validation with each test fold including all six dialect groups, maximizing state coverage, and restricting to one user per location. Each fold is split approximately into 80% training, 10% validation, and 10% testing. In the test set, no speaker augmentation is applied, and the minimum utterance duration is set to two seconds. Due to resource constraints, only the model `mms_coord_a` is evaluated with full 10-fold cross-validation; all other models are evaluated on the first fold only.

### 4.1. Classification and regression results

A model classifying by random chance, representing a uniform selection among all locations, achieves an accuracy of  $\frac{1}{108} \approx 0.93\%$ . All models perform above this level. The geographic center of all locations in the dataset is calculated as  $47.5672^\circ$  latitude and  $13.5636^\circ$  longitude, with an average distance error of 146.41 km to all locations. Only ECAPA-TDNN without speaker augmentation performs worse than random in terms of distance error (153.3 km). Speaker augmentation leads to statistically significant improvements of +4.2% in location accuracy for the XLSR model (`xlsr_loc` vs. `xlsr_loc_nospkaug`, Wilcoxon signed-rank test,  $p < 10^{-35}$ ) and +3.9% for the ECAPA model (`ecapa_loc` vs. `ecapa_loc_nospkaug`,  $p < 10^{-25}$ ).

Predicting coordinates in a regression task reduces the average distance error for all models except ECAPA-TDNN (which exhibits the highest error). Mean prediction error is reduced by 30.1 km with the XLSR model and by 42.2 km with the MMS-LID model, compared to classification on the location level. The best-predicted location, Innervillgraten (Tyr), has a 1.7 km average distance error, while Prellenkirchen (L. Aut)

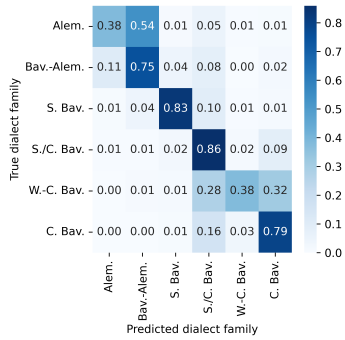


Figure 2: Confusion matrix of dialect groups (*xlsr\_grp\_a*). Accuracies are normalized per row.

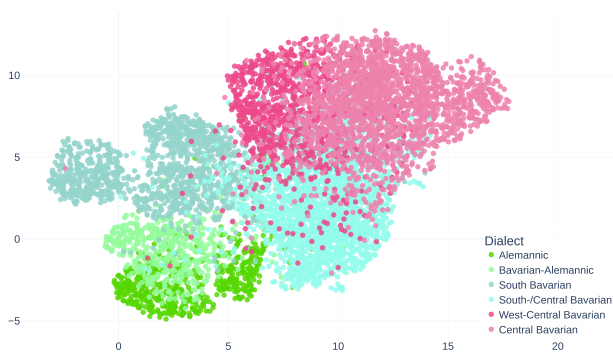


Figure 3: Embeddings of dialect groups (*xlsr\_grp\_a*) projected into two-dimensional space using UMAP.

has the highest at 152.4 km. A 10-fold cross-validation of the MMS-LID model with regression and spectral augmentation (*mms\_coord\_a*) yields a mean distance error of 65.3 km with accuracies of 9.4% at the location level, 65.0% at the dialect group level, and 59.4% at the state level, indicating fold one represents overall performance.

Figure 2 depicts the dialect group confusion matrix: *S./C. Bav.* has the highest accuracy (86%), followed by *S. Bav.* (83%). While *Alem.* is often confused with *Bav.-Alem.* (54%), *W.-C. Bav.* is sometimes misclassified as *C. Bav.* (32%) or *S./C. Bav.* (28%). Confusions mainly occur between neighboring groups; the geographically most distant groups, *Alem.* and *C. Bav.*, are rarely confused ( $\leq 1\%$ ). Trained directly on group labels, the best performance is achieved by the *xlsr\_grp\_a* model (72.7%) trained with spectral augmentation, which improves accuracy by 1.4% over *xlsr\_grp* (71.3%,  $p < 0.005$ ). The *mms\_grp* model achieves an accuracy of 65.9% and performs significantly worse than both *xlsr\_grp* (-5.4%,  $p < 10^{-20}$ ) and *xlsr\_grp\_a* (-6.8%,  $p < 10^{-33}$ ).

Figure 3 displays the embeddings of the dialect groups, revealing that geographically adjacent dialect groups are also clustered closely in the embedding space. Accuracies for federal state classification are depicted in Table 3. The lower accuracies at the state level, compared to the dialect group level, are likely because of the coexistence of multiple dialect groups within individual states. *Tyr* demonstrates the highest performance, while *Bgl* has the lowest due to limited data, small geographic size, multiple bordering states, and the presence of two dialect groups. Its location – bordering both *L. Aut* and *Car* – likely causes confusion with these states (26% and 46%, respec-

Table 3: Classification accuracy by federal states (*xlsr\_state*)

	Bgl	Car	L. Aut	U. Aut	Sbg	Sty	Tyr	Vbg	Vie
	12%	76%	74%	62%	51%	51%	96%	69%	65%

tively). *Sbg* is most often confused with *U. Aut* (26%), possibly because one test sample lies near this border. *Vie* has the fewest data points among all federal states and is frequently misclassified as *U. Aut* (34%), despite being entirely surrounded by *L. Aut*. Limitations of this work are the absence of data from capital cities (except *Vie*) and the bias toward specific demographic groups, which might limit the generalizability to other social or professional groups.

## 4.2. Explainability results

Figure 4 illustrates the most frequently occurring attribution peaks, with phones such as [n] (780), [o] (603), [i] (520), and [ə] (425) being the most prominent. The realization of [ə] varies considerably across dialects, making it a key feature for phonetic analysis and an important cue for model classification. The diphthong [ai] has a greater importance in *S./C. Bav.* than in other dialects. Furthermore, the specific importance of the phoneme /r/ in classifying the *Alem.* (and *Bav.-Alem.*) group could be explained by its realization in this dialect [34]. Knowing the timestamps of highest attribution peaks enables extraction and playback of trimmed audio segments around those peaks. Since certain questions elicit dialectal phones, the higher average kurtosis in the corresponding utterances (158.7 vs. 149.0 for all other questions) supports the dialectological validity of IG-based explainability, indicating that more pronounced attribution peaks occur when specific phones are present.

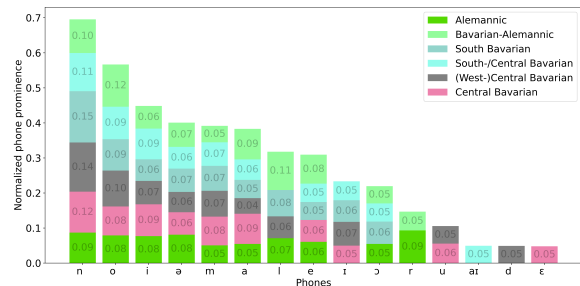


Figure 4: Top 10 phones of each dialect group, normalized by their frequency of occurrence within each dialect group.

## 5. Conclusion

This study compares dialect classification with geographic coordinate prediction, demonstrating that the latter reduces the average distance error. It provides a detailed evaluation of dialect identification solely from audio, leveraging state-of-the-art models such as XLSR and MMS-LID. While XLSR performs best when directly trained on dialect group labels (72.7% accuracy), MMS-LID achieves the lowest distance error (66.7 km) in the regression task. Additionally, attribution analysis on the raw waveform highlights the most influential phonetic features for dialect categorization. Future research could integrate additional datasets to test the models on unseen locations, but especially to enhance granularity and further improve classification accuracy.

## 6. References

- [1] M. Hornung and F. Roitinger, *Die österreichischen Mundarten. Eine Einführung - Austrian dialects. An introduction.* Wien: öbv&hpt, 2000, 160 pages.
- [2] S. Moosmüller, *Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck - Standard and dialect in Austria - A sociophonological study on teasing apart standard and dialect in Vienna, Graz, Salzburg, and Innsbruck.* Wien: Böhlau, 1991, 212 pages.
- [3] S. Moosmüller, C. Schmid, and J. Brandstätter, “Standard Austrian German,” *Journal of the International Phonetic Association*, vol. 45, no. 3, pp. 339–348, 2015.
- [4] M. Stadtschnitzer, C. A. Schmidt, and D. Stein, “Towards a localised German automatic speech recognition,” in *ITG Symposium on Speech Communication*, 2014.
- [5] J. Linke, B. C. Geiger, G. Kubin, and B. Schuppler, “What’s so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures,” *Comput. Speech Lang.*, vol. 90, p. 101738, 2024.
- [6] R. Xie, O. Ahia, Y. Tsvetkov, and A. Anastasopoulos, “Extracting lexical features from dialects via interpretable dialect classifiers,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, vol. 2. Association for Computational Linguistics (ACL), 2024, p. 54 – 69.
- [7] A. Akman and B. W. Schuller, “Audio explainable artificial intelligence: A review,” *Intelligent Computing*, vol. 3, Jan. 2024.
- [8] Z. Al-Jumaili, T. Bassiouny, A. Alanezi, W. Khan, D. Al-Jumeily, and A. J. Hussain, *Classification of Spoken English Accents Using Deep Learning and Speech Analysis.* Springer International Publishing, 2022, pp. 277–287.
- [9] J. Zuluaga-Gomez, S. Ahmed, D. Visockas, and C. Subakan, “CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice,” in *Proc. Interspeech 2023*, 2023, pp. 5291–5295.
- [10] J. Dobbriner and O. Jokisch, *Towards a Dialect Classification in German Speech Samples.* Springer International Publishing, 2019, pp. 64–74.
- [11] H. Wagner, “Austrian dialect classification using machine learning,” Master’s thesis, FH Hagenberg, Interactive Media, Hagenberg, Austria, June 2019.
- [12] P. C. Vergeiner, “Dialect Classification and Everyday Culture: A Case Study from Austria,” *Languages*, vol. 10, no. 2, 2025.
- [13] T. Kisler and F. Schiel, “Towards a speaker localization from spontaneous speech: north-south classification for speakers of contemporary German,” in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, A. Berton, U. Haiber, and W. Minker, Eds. TUDpress, Dresden, 2018, pp. 200–207.
- [14] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapschkin, and W. Samek, “Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark,” *Journal of the Franklin Institute*, vol. 361, no. 1, p. 418 – 428, 2024.
- [15] F. Iqbal, Z. S. Syed, M. S. S. Syed, and A. S. Syed, “An Explainable AI Approach to Speech-Based Alzheimer’s Dementia Screening,” in *SMM24, Workshop on Speech, Music and Mind 2024*, ser. smm.2024. ISCA, 2024, pp. 11–15.
- [16] J. Parekh, S. Parekh, P. Mozharovskiy, F. d’Alché-Buc, and G. Richard, “Listen to Interpret: Post-hoc Interpretability for Audio Networks with NMF,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 35 270–35 283.
- [17] V. Haunschmid, E. Manilow, and G. Widmer, “audioLIME: Listenable Explanations Using Source Separation,” 13th International Workshop on Machine Learning and Music, 2020.
- [18] B. Desplanques, J. Thienpondt, and K. Demuyne, “ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.
- [19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.
- [20] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [21] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 3319–3328.
- [22] P. Wiesinger, “Die Einteilung der deutschen Dialekte,” in *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, W. Besch, U. Knoop, W. Putschke, and H. E. Wiegand, Eds. de Gruyter, 1983, vol. 2, pp. 807–900.
- [23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, p. 4218 – 4222.
- [24] M. Pucher, F. Neubarth, V. Strom, S. Moosmüller, G. Hofer, C. Kranzler, G. Schuchmann, and D. Schabus, “Resources for speech synthesis of Viennese varieties,” in *LREC 2010 - 7th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2010, p. 105 – 108.
- [25] D. Schabus, M. Pucher, and G. Hofer, “Joint audiovisual Hidden semi-Markov Model-based speech synthesis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347, 2014.
- [26] H. Bredin, “pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe,” in *Proc. Interspeech 2023*, vol. 2023-August. International Speech Communication Association, 2023, p. 1983 – 1987.
- [27] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, “Data augmentation improves recognition of foreign accented speech,” in *Proc. Interspeech 2018*, 2018.
- [28] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.
- [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [30] J. Li, W. Tu, and L. Xiao, “Freevc: Towards high-quality text-free one-shot voice conversion,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2023-June. Institute of Electrical and Electronics Engineers Inc., 2023.
- [31] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for pytorch,” 2020.
- [32] A. Savitzky and M. J. E. Golay, “Smoothing and differentiation of data by simplified least squares procedures,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [33] Q. Xu, A. Baevski, and M. Auli, “Simple and effective zero-shot cross-lingual phoneme recognition,” in *Proc. Interspeech 2022*, vol. 2022-September. International Speech Communication Association, 2022, p. 2113 – 2117.
- [34] A. Leemann, S. Schmid, D. Studer-Joho, and M.-J. Kolly, “Regional variation of /r/ in Swiss German dialects,” in *Proc. Interspeech 2018*, 2018, pp. 2738–2742.